INFORMATION AND CONTROL 9, 279–284 (1966)

Necessary and Sufficient Conditions on Conditional Probabilities to Maximize Entropy*

CHARLES E. RADKE

International Business Machines Corporation, Systems Development Division, Endicott, New York

Shannon (1948) showed that, by a proper choice of the conditional probabilities of the symbols in a discrete noiseless channel of the finite-state type (which possesses certain inherent constraints), the entropy of symbols on such a channel could be maximized. To date these known sufficient conditions on the conditional probabilities have been assumed to also be the necessary conditions. In this brief paper it is shown that the sufficient conditions as originally stated by Shannon are indeed also the necessary conditions for the range of symbol lengths which are of interest.

Information sources which can be represented by finite state diagrams have been of interest at least since the paper of Shannon (1948). Recent interest in transformations for coding purposes of one finite-state device into another device which possesses symbols of possibly different lengths has been shown, for example, by Karp (1961) and Neumann (1962). Concerning the efficiency of the encoding (transformation) of such finite-state devices, where one finite-state device defines an information source and one a channel, Shannon (1948, 1949) in his Theorem 8 states certain conditions on the conditional probabilities. These stated conditions on the conditional probabilities (given the state) of symbols occurring, as determined by the source, are such that the entropy H of the symbols on the channel is maximized and equal to the channel capacity C when the average symbol length l_{ave} is equal to unity.

To date no statement concerning the necessary conditions on the conditional probabilities has been made or proved. Necessary and sufficient conditions are proved here on the conditional probabilities in order to maximize the ratio $\eta_c = (H/l_{\text{ave}})/C$; from this result the necessity of the conditions stated in Shannon's theorem immediately follows. In

* This work was supported in part by the National Science Foundation.

order to obtain the proof concerning the conditions on the *conditional probabilities*, the necessary and sufficient conditions on the *symbol lengths* for maximum η_c are first determined. It is then shown that the conditions on the conditional probabilities must be identical to those proven on the symbol lengths.

TERMINOLOGY

A system of constraints on sequences of symbols which serves as a channel can be expressed conveniently in terms of a state diagram, i.e., a directed graph in which vertices are denoted as states and arcs as symbols of various lengths. Such channels may be referred to as finite-state channels (Radke, 1964). With such a channel, a set of states, say n states, may be associated. Let m_{ij}^{v} denote the vth symbol from state s_i to s_j, l_{ij}^{v} the length associated with the symbol. Assume that the statistical properties of the information source are completely known, that encoding has taken place, and messages are now represented by strings of symbols. Then, given that the channel is in state s_i, the conditional probability p_{ij}^{v} of this symbol can be determined. The information source also determines the probability P_i of the channel being in the ith state. Let H be the entropy of the symbol on such a channel,

$$H = -\sum_{i} \sum_{j} \sum_{v} P_i \, p_{ij}^{v} \log_x p_{ij}^{v},$$

where

$$\sum_{j} \sum_{v} p_{ij}^{v} = 1 \quad \text{and} \quad \sum_{i} P_i = 1;$$

let l_{ave} be the average symbol length,

$$l_{\text{ave}} = \sum_{i} \sum_{j} \sum_{v} P_i \, p_{ij}^{v} l_{ij}^{v};$$

and C be the channel capacity in x-ary digits per second, $C = \log_x \omega_r$. The channel capacity is entirely independent of all properties of the information source. Define

$$A \triangleq [\textstyle\sum \omega^{-l_{ij}^{v}} - \delta_{ij}];$$

then from Theorem 1 in Shannon (1949) ω_r is the maximum positive solution to

$$|A| = 0. \tag{1}$$

PROBLEM

Shannon (1948, 1949) in his Theorem 8 states that if

$$p_{ij}^{\ v} = \frac{\beta_j}{\beta_i}\,\omega_r^{-l_{ij}v} \tag{2}$$

and $\sum_j \sum_v p_{ij}^{\ v} = 1$, then H is maximized and equal to C when $l_{\text{ave}} = 1$. It is required to prove that (2) is necessary as well as sufficient to maximize the ratio $\eta_c = (H/l_{\text{ave}})/C$. Further, it is desired to show that such conditions are necessary and sufficient for any $l_{\text{ave}} > 0$.

SOLUTION TO PROBLEM

In order for the channel to have meaning it must possess a strongly connected property, that is, any channel state must be able to be reached from any other channel state. Then the matrix A can be shown to have three properties which are stated as lemmas below.

LEMMA 1 (Radke, 1964). *All minors of order $n - 1$ of the matrix A are nonzero; i.e., no minor of order $n - 1$ of A is identically zero.*

Proof: The lemma follows from the strongly connected property of finite-state channels.

LEMMA 2 (Radke, 1964). *Given A, then $A_{ij}/A_{ii} = A_{kj}/A_{ki}$ for all i, j, and k, where A_{ij} are cofactors of A.*

Proof: The lemma follows from the fact that the adjoint of any singular matrix has all its minors of order two identically zero.

LEMMA 3. *Let A_{ii} and A_{ij} be cofactors of A; then*

$$\prod_{i=1}^{n} \prod_{j=1}^{n} (A_{ii}/A_{ij}) = 1\,.$$

Proof: The proof is by induction through the use of Lemma 2.

Shannon's Theorem 8 states sufficient conditions on the set of conditional probabilities $\{p_{ij}^{\ v}\}$ of the symbols and assumes that the set of lengths $\{l_{ij}^{\ v}\}$ is known. The converse is now assumed. Assume that the set of conditional probabilities is known, i.e., an information source is given, but that the set of symbol lengths is unknown, i.e., the channel is not fully described. Let us then find the necessary and sufficient conditions on the set of symbol lengths.

THEOREM 1. *A necessary and sufficient condition such that η_c is maximized and equal to 1 is that the set of symbol lengths $l_{ij}^{\ v}$ must satisfy*

$$p_{ij}^{\ v} = (\beta_j/\beta_i)\omega_r^{-l_{ij}v}, \tag{3}$$

where β_i and β_j are respectively the (k, i)th and (k, j)th cofactor of the matrix A for any k.

Proof: The proof reduces to finding the stationary points for η_c under the constraint (1). However, these stationary points are the same as for the product $l_{\text{ave}} \times C$ under the same constraint. Further, all possible maximum points for the product must be one or more of the following:

(1) boundary points
(2) stationary points
  (a) solution to the resultant Lagrange equations
  (b) singular points on the locus of the constraint.

Let $y_{ij}^{v} = \omega_r^{-l_{ij}^{v}}$. The restrictions on ω_r and l_{ij}^{v} are $1 \leqq \omega_r < \infty$ and $< l_{ij}^{v} < \infty$; let the restrictions on y_{ij}^{v} be

$$0 \leqq y_{ij}^{v} \leqq 1. \tag{4}$$

The restriction (4) can be written as $y_{ij}^{v} = \text{Sin}^2 \theta_{ij}^{v}$. Therefore the expression required in order to find the stationary points of $l_{\text{ave}} \times C$ under constraints, including boundary constraints, is

$$J = -\sum_i \sum_j \sum_v P_i p_{ij}^{v} \log_x y_{ij}^{v} + \lambda \mid A \mid + \sum_i \sum_j \sum_v u_{ij}^{v}(y_{ij}^{v} - \text{Sin}^2\theta_{ij}^{v}),$$

where the λ and u_{ij}^{v} are Lagrangian multipliers. The resultant Lagrange equations are

$$\begin{aligned} \partial J/\partial y_{ij}^{v} &= -P_i p_{ij}^{v}(1/y_{ij}^{v}) \log_x e + \lambda A_{ij} + u_{ij}^{v} = 0 \\ \partial J/\partial \theta_{ij}^{v} &= -2u_{ij}^{v} \text{ Sin } \theta_{ij}^{v} \text{ Cos } \theta_{ij}^{v} = 0, \end{aligned}$$

where A_{ij} is the (i, j)th cofactor of A. To determine the stationary points not on the boundary assume $u_{ij}^{v} = 0$ for all i, j, and v.

The solution of the Lagrange equations produces a unique multiplier $\lambda = (P_i \log_x e)/A_{ii}$. From Lemma 1, $A_{ij} \neq 0$ for all i and j, and hence $p_{ij}^{v} = (A_{ij}/A_{ii})y_{ij}^{v}$. From Lemma 2 $A_{qp}/A_{qi} = A_{mp}/A_{mi}$ for all m, q, and since m, q are arbitrary, let $A_{qj}/A_{qi} = \beta_j/\beta_i$ for any q and all i and j. Therefore, the stationary points are defined by (3).

The boundary values (when at least one $u_{ij}^{v} \neq 0$) must now be evaluated. The function with which we are working is a composite of logarithmic functions and thus causes difficulties in analyzing the boundary values. Observe that at $y_{ij}^{v} = 0$ the first partial derivative $\partial J/\partial y_{ij}^{v}$ does not exist but does exist in the interval $0 < y_{ij}^{v} \leqq 1$. Further, the function η_c can have one of two domains which satisfy the constraints

$|A| = 0$ and $0 \leqq y_{ij}^{v} \leqq 1$. Either the domain of η_c is the point—all $y_{ij}^{v} = 1$—or the domain is the set of points which satisfy $|A| = 0$ and $0 \leqq y_{ij}^{v} \leqq 1$ and not all $y_{ij}^{v} = 1$.

If any $y_{ij}^{v} = 1$, then in order to obey the constraint the channel must be such that all y_{ij}^{v} which appear in the sums must be either 0 or 1. If the y_{ij}^{v} of zero value are neglected, only one path may exist from one state to another. For example, in a three-state device, if $y_{13}^{1} = 1$, then $y_{32}^{1} = 1$, $y_{21}^{1} = 1$, and all other y_{ij}^{v} must either not exist (not appear in the sums) or be zero.

If there are no y_{ij}^{v} of zero value and at least one $y_{ij}^{v} = 1$, *then all* $y_{ij}^{v} = 1$. Then, to preserve the property $\sum_j \sum_v p_{ij}^{v} = 1$, the corresponding $p_{ij}^{v} = 1$. Therefore, for the condition when all $y_{ij}^{v} = 1$, $\eta_c \equiv 1$ and the domain of η_c is a single point. Observe that whenever any $y_{ij}^{v} = 1$, condition (3) is not violated.

However, if any $y_{ij}^{v} = 0$ (this particular y_{ij}^{v} exists in the sum), there is a nonzero probability that the corresponding symbol m_{ij}^{v} will occur, i.e., $p_{ij}^{v} > 0$. The condition that some $p_{ij}^{v} = 1$ is not prevented; however, all p_{ij}^{v} cannot be unity. Since the numerator of η_c is finite and nonzero ($H \neq 0$) and the denominator is very large without bound, for any $y_{ij}^{v} = 0$, $\eta_c = 0$. Definitely for some $y_{ij}^{v} = 0$ the point is not maximum; in fact, since $\eta_c \geqq 0$, this point must be a minimum although the derivative of η_c does not exist at $y_{ij}^{v} = 0$. Since the function η_c is continuous over the defined domain when not all $y_{ij}^{v} = 1$ and goes to zero for any $y_{ij}^{v} = 0$, the maximum points of η_c must appear as stationary points in the range $0 < y_{ij}^{v} < 1$.

That (3) results in $\eta_c = 1$ can be readily shown with the aid of Lemma 2.

Let us now return to the condition in which the set of lengths is assumed known and necessary and sufficient conditions are to be found on the set of conditional probabilities.

THEOREM 2. *The ratio* η_c *for all finite-state channels is maximized and equal to* 1 *if and only if* p_{ij}^{v} *for all* i, j, *and* v *satisfy condition* (3).

Proof: Sufficiency is from Theorem 8 of Shannon (1949). Necessity follows from Theorem 1 since the mapping described by (3) is the only mapping which will give $\eta_c = 1$.

COROLLARY 1. *The entropy* H *is maximized and equal to* $C(l_{\text{ave}} = 1)$ *if and only if the* p_{ij}^{v} *satisfy* (3).

Proof: The proof follows from Theorem 2 since the product $l_{\text{ave}} \times C$ is known.

An additional property of the matrix A can be shown when $\eta_c = 1$.

THEOREM 3. *For* $\eta_c = 1$ *the matrix* A *is similar to the matrix*

$$P \triangleq [\sum_{\nu} p_{ij}^{\nu} - \delta_{ij}].$$

Proof: Let

$$T = \begin{bmatrix} 1/\beta_1 & 0 & \cdots & 0 \\ 0 & 1/\beta_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1/\beta_n \end{bmatrix} \quad \text{and} \quad T^{-1} = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \beta_n \end{bmatrix};$$

then $TAT^{-1} = P$.

Besides the result of the necessary and sufficient conditions on the conditional probabilities (or lengths) of the symbols, Theorems 1 and 2 show that the capacity of a finite-state channel as defined by Shannon is indeed the maximum rate of transmission of information possible over such a channel.[1]

ACKNOWLEDGMENTS

The author is grateful to the referee for his constructive criticisms.

RECEIVED: January 13, 1965
REVISED: January 11, 1966

REFERENCES

BLACKWELL, D., BREIMAN, L., AND THOMASIAN, A. J., (1958), Proof of Shannon's transmission theorem for finite-state indecomposable channels. *Ann. Math. Statist.* **20,** 1209–1220.

KARP, R. M., (1961), Minimum-redundancy coding for the discrete noiseless channel. *IRE Trans. Inform. Theory,* **IT-7,** 27–38.

NEUMANN, P. G., (1962), Efficient error-limiting variable-length codes. *IRE Trans. Inform. Theory,* **IT-8,** 292–304.

RADKE, C. E., (1964), "Generalized Aspects of Noiseless, Discrete Finite-State Channels." Systems Research Center Report No. 48-A-64-16 (Case Institute of Technology, Cleveland, Ohio).

SHANNON, C. E., (1948), The mathematical theory of communications. *Bell System Tech. J.* **27,** 379–423.

SHANNON, C. E., AND WEAVER, W., (1949), "The Mathematical Theory of Communications." Univ. of Illinois Press, Urbana, Illinois (ninth printing, 1962).

[1] Another proof of this last result is given in a paper by Blackwell, Breiman, and Thomasian (1958).